NOTE

USING *DAUBERT* TO EVALUATE EVIDENCE-BASED SENTENCING

Charlotte Hopkinson†

† Charlotte Hopkinson is a student in the dual degree Cornell Law School and Université Paris I Panthéon-Sorbonne program and will graduate in 2019 with a JD and Master en Droit. Special thanks to all those who were willing to discuss my Note with me, Joe Margulies and John Zipp. John Zipp was invaluable in interpreting the methods and results published by the Pennsylvania Commission on Sentencing. This Note was inspired by my Science and the Law Class at Auburn Correctional Facility, which was taught through the Cornell Prison Education Program in Fall 2016. Student-led class discussions helped me consider various legal and policy viewpoints.

INTRODUCTION

Jack and Jill went up the hill,
to steal a pail of water,
Both were caught and sentenced to jail,
But Jack came out two years later.

Why? Assume that both Jack and Jill's cases are identical in facts, procedure, jury composition, and verdict. The only relevant difference is that Jack is a man and Jill is a woman. Statistically, men are more likely to recidivate than women,[1] and under a sentencing system called evidence-based sentencing, the judge agrees that Jack is more likely to commit another crime due to his gender. The solution to this prediction? Jack will spend more time in prison, simply because he is a man.

Evidence-based sentencing is part of a growing trend of using actuarial risk assessments in the criminal justice system.[2] Evidence-based sentencing relies on a set of factors, some of which are relevant to the crime committed but most of which are based on immutable characteristics outside of the defendant's control, to predict the probability that a defendant will recidivate.[3] Notwithstanding the constitutionally problematic nature of relying on certain factors, such as sex,[4] the scientific validity of such methods is questionable. Although evidentiary rules do not apply to sentencing, this Note argues that the admissibility of such evidence in general, and specifi-

---

1 Sonja B. Starr, *Evidence-Based Sentencing and the Scientific Rationalization of Discrimination*, 66 STAN. L. REV. 803, 863 (2014) ("To give a simple illustration, if a sentence is based on crime severity plus gender, and these factors together produce a ten-year sentence for a male when an otherwise-identical woman would have received seven years, male gender is not solely responsible for the sentence; crime severity establishes the baseline of seven years. But male gender is solely responsible for the extra three years.").

2 Dawinder S. Sidhu, *Moneyball Sentencing*, 56 B.C. L. REV. 671, 673 (2015) ("More and more, courts today are adopting the use of risk-assessment tools in sentencing. These risk-assessment tools take information on recidivism rates for groups and use them to estimate the risk of recidivism for individuals possessing those same group characteristics.").

3 The topic of predictive factors of dangerousness recently arose in oral arguments in front of the Supreme Court in *Buck v. Davis*. Transcript of Oral Argument at 31, Buck v. Davis, 137 S. Ct. 759 (2016) (No. 15-8049).

4 The author of this Note would like to point out that actuarial instruments and researchers tend to use gender and sex interchangeably. These two words, however, have two different meanings. For the purpose of this Note, sex is biologically defined, whereas gender is a societal construct. The author also uses the singular "they" throughout this Note.

cally Pennsylvania's new sentencing scheme, fails the *Daubert* framework.

Part I of this Note provides a background into the characteristics and history of different types of evidence-based sentencing and addresses the arguments for and against its use. Part II describes the "future dangerousness" standard used in death penalty sentencing and civil commitments of sexually violent persons as an analogy to evidence-based sentencing and argues that evidence-based sentencing must withstand some type of evidentiary gatekeeping test. Part III subjects evidence-based sentencing to the *Daubert* test and applies this test broadly to evidence-based sentencing generally, as well as specifically to Pennsylvania's Risk Assessment Project.[5] Part IV uses Federal Rule of Evidence 403 to further explain why evidence-based sentencing should not be used, particularly for models that include sex and age. Part V briefly addresses certain penological theories that evidence-based sentencing reflects and questions whether evidence-based sentencing is the appropriate solution.

I
BACKGROUND TO EVIDENCE-BASED SENTENCING

Evidence-based sentencing focuses on predicting an individual's recidivism risk based on empirical research.[6] Evidence-based sentencing is a type of risk assessment, or actuarial method, that relies on a large dataset to evaluate the "statistical correlations between a group trait and that group's criminal offending rate" as opposed to a clinical evaluation.[7] Empirical research attributes recidivism to a wide range of factors, including criminal history, sex, age, marital status, employment, education, parental convictions, family members who have been crime victims, high school grades, chances of finding work above minimum wage, dependence on social as-

---

5 *See* PA. COMM'N ON SENT'G, OVERVIEW OF THE RISK ASSESSMENT INSTRUMENT, http://pcs.la.psu.edu/publications-and-research/research-and-evaluation-reports/risk-assessment [https://perma.cc/5FUC-GVX2].

6 Starr, *supra* note 1, at 805.

7 Bernard H. Harcourt, *Against Prediction: Sentencing, Policing, and Punishing in an Actuarial Age* 11 (U. of Chi. Pub. Law & Legal Theory Working Paper Series, Paper No. 94, 2005); *see* Jordan M. Hyatt, Mark H. Bergstrom & Steven L. Chanenson, *Follow the Evidence: Integrate Risk Assessment into Sentencing*, 23 FED. SENT'G REP. 266, 266 (2011) [hereinafter *Follow the Evidence*] (noting that actuarial risk assessments rely on static variables whereas clinical risk assessments rely on dynamic variables).

sistance, finances, and neighborhood of residence.[8] By using an impersonal, data-driven method to determine the recidivism risk, many actors in the criminal justice system believe that they are simultaneously increasing fairness in discretionary sentencing and reducing the overpopulation of prisons by diverting low-risk offenders from prison.[9]

Although probabilistic reasoning is found throughout the criminal justice system, such as a police officer's determination that there is sufficient cause for a search, forensic testimony concerning a DNA "match," or the legal standard for criminal guilt itself ("beyond a reasonable doubt"), these examples of probabilistic reasoning are not based on statistical correlations of a group trait, but rather on a situational case-by-case analysis.[10] This new focus on evidence-based sentencing changes the relationship between group data and individual predictions in the criminal justice system and results in serious consequences (loss of liberty). An increasing amount of states are adopting evidence-based sentencing guidelines,[11] and the American Law Institute's Model Penal Code[12] draft incorpo-

---

8 These factors differ between states and assessment methods. Researchers have identified hundreds of factors relevant to sentencing but are trying to narrow them down to the few factors that are the most predictive. *See* Starr, *supra* note 1, at 805, 811–13; Sonja B. Starr, *Sentencing, by the Numbers*, N.Y. TIMES (Aug. 10, 2014), http://www.nytimes.com/2014/08/11/opinion/sentencing-by-the-numbers.html [https://perma.cc/EML7-TRAY]; *see also* J.C. Oleson, *Risk in Sentencing: Constitutionally Suspect Variables and Evidence-Based Sentencing*, 64 S.M.U. L. REV. 1329, 1350–51 (2011) (distinguishing "having criminal companions," "antisocial personality," "criminogenic needs," "adult criminal history," and "race" as strongly significant predictors of recidivism versus "substance abuse," "family structure," "intellectual functioning," "family criminality," "gender," "socio-economic status of origin," and "personal distress" as weaker, but still significant, predictors). For a list of 125 factors, see PA. COMM'N ON SENT'G, INTERIM REPORT 1: REVIEW OF FACTORS USED IN RISK ASSESSMENT INSTRUMENTS 13–17 (2011).

9 Starr, *supra* note 8; *see* Julia Angwin, Jeff Larson, Surya Mattu & Lauren Kirchner, *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [https://perma.cc/8ZPQ-9F92]; *Follow the Evidence*, *supra* note 7, at 267 (arguing that the sentences would be facially disparate, but individualized and thus fair). The use of evidence-based models is also used in bail determinations. Shaila Dewan, *Judges Replacing Conjecture with Formula for Bail*, N.Y. TIMES (June 26, 2015), http://www.nytimes.com/2015/06/27/us/turning-the-granting-of-bail-into-a-science.html [https://perma.cc/5X6J-5KD8].

10 Harcourt, *supra* note 7, at 11.

11 *See* NANCY LAVIGNE ET AL., URBAN INST., JUSTICE REINVESTMENT INITIATIVE STATE ASSESSMENT REPORT 39–40 (2014), https://www.urban.org/sites/default/files/publication/22211/412994-Justice-Reinvestment-Initiative-State-Assessment-Report.PDF [https://perma.cc/74XL-3N3K]; *see also* Starr, *supra* note 8 (noting that "this practice has rapidly expanded much more recently").

12 MODEL PENAL CODE: SENTENCING app. A at 133, 135 (AM. LAW INST., Discussion Draft No. 4, 2012).

rates a recommendation to use actuarial instruments.[13] The use of predictive instruments to estimate recidivism originated with discretionary parole (which, ironically, has been abolished in sixteen states)[14] but has expanded to more than twenty states' current sentencing practices.[15] While Pennsylvania is debuting its own predictive instrument,[16] other states use generic actuarial tests that are only sometimes calibrated to state-specific populations. Two of the most common generic actuarial tests are the Level of Services Inventory-Revised (LSI-R) and the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS).[17] Like many of the state-specific tests, these commercial tests include objective and subjective variables that are unrelated to the defendant's crime.[18]

Both legal and scholarly advocates for evidence-based sentencing believe in its ability to simultaneously hold offenders accountable, protect the public, and reduce incarceration (and associated financial costs and social harms).[19] Evidence-based sentencing purportedly offers a "scientific" way to systematically differentiate low-risk and high-risk offenders. Advocates argue that judges already engage in predictions during sentencing—therefore, evidence-based sentencing will guide judges with scientific feedback and increase sentencing transparency.[20] The purported benefits of this transparency extend

---

13 *See* Starr, *supra* note 1, at 805.

14 TIMOTHY A. HUGHES ET AL., U.S. DEP'T OF JUSTICE, BUREAU OF JUSTICE STATISTICS, TRENDS IN STATE PAROLE, 1990–2000 (2001). In 1994, Virginia abolished parole and established a risk assessment instrument for reducing incarceration rates. David A. Soulé & Stacy S. Najaka, *Paving a Path to Informed Sentencing Decisions*, 25 FED. SENT'G REP. 181, 181 (2013).

15 Starr, *supra* note 1, at 809 ("A review of case law, sentencing commission websites, and relevant legislation indicates that at least twenty states' courts are now, in some or all cases, incorporating actuarial risk assessments into the determination of the defendant's sentence.").

16 Logan Koepke, *Pennsylvania Will Vary Jail Terms for the Same Crime, Based on Where You Live*, EQUAL FUTURE (Sept. 16, 2015), https://www.equalfuture.us/2015/09/16/pennsylvania-will-vary-jail-terms-for-the-same-crime-based-on-where-you-live/ [https://perma.cc/Z7HU-UQMN].

17 Starr, *supra* note 1, at 812.

18 *Id.* at 812–13.

19 *See id.* at 816; Soulé & Najaka, *supra* note 14, at 181; *see also Follow the Evidence*, *supra* note 7, at 266 ("The use of risk assessment at sentencing underscores an overall shift in the purposes of sentencing, moving from a backward-looking retributive approach with a focus on uniformity, proportionality, and reduction of unwarranted disparity to an approach that also incorporates a formalized, forward-looking utilitarian goal.").

20 *See* Lynn S. Branham, *Follow the Leader: The Advisability and Propriety of Considering Cost and Recidivism Data at Sentencing*, 24 FED. SENT'G REP. 169, 171 (2012).

to the general public who will perceive the legal system as more legitimate, thus causing them to comply with more laws.[21]

Evidence-based sentencing has also come under scrutiny and criticism.[22] Scholars note that because evidence-based sentencing focuses on uncontrollable characteristics of the defendant rather than the crime at hand, it distorts penological goals, particularly the retributive and rehabilitative goals.[23] Other critics, particularly in the media, analogize evidence-based sentencing to the movie *Minority Report* in which the government punishes crimes before they happen, because a person's prison sentence is partially based on their future likelihood to recidivate.[24] Furthermore, the constitutionality of the evidence-based sentencing is at question.[25] The use of demographic factors including sex and socioeconomic status raises constitutional issues such as equal protection and the rights of indigent defendants.[26] Although these scholars support reducing incarceration rates, they predict that doing so by considering demographic factors risks further exacerbating the inequities of the criminal justice system.[27] These critiques are not limited to academics. Although some courts have apprehensively embraced evidence-based sentencing,[28] other actors

---

21 *See* Tracey L. Meares, *Three Objections to the Use of Empiricism in Criminal Law and Procedure—And Three Answers*, 2002 U. ILL. L. REV. 851, 866 (2002).

22 *See* Starr, *supra* note 1, at 817 ("Although most of the literature on EBS is positive, or even celebratory, a few scholars have criticized it.").

23 *See* Harcourt, *supra* note 7, at 31–32 ("We have come to associate the prediction of future criminality with just punishment."); *see also* Starr, *supra* note 1, at 817 ("[P]rediction instruments contravene punishment theory, because punishment turns on *who the defendant is* . . . .").

24 Jared Greenhouse, *Pennsylvania Wants to Use Science in Criminal Sentencing*, HUFFINGTON POST (Aug. 11, 2015, 11:54 AM), http://www.huffingtonpost.com/entry/science-criminal-sentencing_us_55c8ed49e4b0923c12bda693 [https://perma.cc/U272-PUS8]; *see* Anna Maria Barry-Jester, Ben Casselman & Dana Goldstein, *The New Science of Sentencing*, THE MARSHALL PROJECT (Aug. 4, 2015, 7:15 AM), [hereinafter *The New Science of Sentencing*] https://www.themarshallproject.org/2015/08/04/the-new-science-of-sentencing#.4mF2KXDVR [https://perma.cc/9465-4MND].

25 *See* Starr, *supra* note 1, at 821–42; Oleson, *supra* note 8, at 1372–88; Shaina D. Massie, Note, *Orange Is the New Equal Protection Violation: How Evidence-Based Sentencing Harms Male Offenders*, 24 WM. & MARY BILL RTS. J. 521, 522 (2015) ("Penological considerations of gender in sentencing are simply incompatible with abstract notions that criminal offenders appear before the court in their individual capacities. More important, the use of gender in evidence-based sentencing violates the concrete promises of equal protection under the law provided by the Constitution.").

26 *See* Starr, *supra* note 1, at 806.

27 *See id.*

28 Danielle Citron, *(Un)Fairness of Risk Scores in Criminal Sentencing*, FORBES (July 13, 2016, 3:26 PM), http://www.forbes.com/sites/daniellecitron/2016/07/13/unfairness-of-risk-scores-in-criminal-sentencing/#46d161f54479 [https://

in the justice system, such as former Attorney General Eric Holder, have warned of the dangers of using immutable characteristics in sentencing decisions.[29]

Both sides of the evidence-based sentencing debate have laudable goals; however, it is difficult, ironically, to predict the outcome of evidence-based sentencing in the criminal justice system. Rather than focus on the hypothetical positive and negative outcomes or the constitutionality of this sentencing practice, this Note seeks to further contribute to academic discussion by exploring the scientific validity of such instruments.

II
SCIENTIFIC VALIDITY OF PREDICTIVE INSTRUMENTS DURING SENTENCING

Much of the legal and academic discussion surrounding the validity and use of predictive instruments is focused on two areas of the law: future dangerousness predictions for capital

---

perma.cc/H4DJ-FQYB] ("[The Wisconsin Supreme Court] made clear its concerns about the accuracy and potential bias of risk scoring systems. The Court recognized that while states like New York have studied the effectiveness and predictive accuracy of COMPAS scores and found them 'satisfactory,' Wisconsin had not completed a statistical validation study for COMPAS for its population.").

29 Attorney General Holder emphasized that he was "concerned that [risk assessments] may inadvertently undermine our efforts to ensure individualized and equal justice. By basing sentencing decisions on static factors and immutable characteristics—like the defendant's education level, socioeconomic background, or neighborhood—they may exacerbate unwarranted and unjust disparities that are already far too common in our criminal justice system and in our society." Eric Holder, Attorney General, Speech at the National Association of Criminal Defense Lawyers 57th Annual Meeting and 13th State Criminal Justice Network Conference (Aug. 1, 2014) (transcript available at https://www.justice.gov/opa/speech/attorney-general-eric-holder-speaks-national-association-criminal-defense-lawyers-57th [https://perma.cc/8J5G-FFKU]); *see also* Letter from Jonathan J. Wroblewski, Director, Office of Policy and Legislation, U.S. Department of Justice, to Hon. Patti B. Saris, Chair, U.S. Sentencing Commission, at 7 (July 29, 2014), https://www.justice.gov/sites/default/files/criminal/legacy/2014/08/01/2014annual-letter-final-072814.pdf [https://perma.cc/WMP8-ZGUV] ("[D]etermin[ing] risk levels based on static, historical offender characteristics such as education level, employment history, family circumstances and demographic information . . . [and] basing criminal sentences, and particularly imprisonment terms, primarily on such data—rather than the crime committed and surrounding circumstances—is a dangerous concept that will become much more concerning over time as other far reaching sociological and personal information unrelated to the crimes at issue are incorporated into risk tools. This phenomenon ultimately raises constitutional questions because of the use of group-based characteristics and suspect classifications in the analytics. Criminal accountability should be primarily about prior bad acts proven by the government before a court of law and not some future bad behavior predicted to occur by a risk assessment instrument.").

cases[30] and for sexually violent person civil commitments.[31] This section reviews the relevant legal and academic discourse and analogizes the use of future dangerousness predictions in these areas to the use of evidence-based predictions at sentencing. Although courts have found the use of future dangerousness predictions in capital cases and civil commitments admissible, many academics argue that a gatekeeping test like *Daubert* would disallow this kind of evidence.[32]

A. Capital Cases

In the sentencing stages of capital cases, experts may testify on a capital defendant's future dangerousness.[33] In *Barefoot v. Estelle*,[34] the United States Supreme Court refused to categorically exclude a psychiatrist's testimony about the defendant's future dangerousness in sentencing, despite the American Psychiatric Association's assertion that these kinds of predictions are wrong *most* of the time.[35] In this case, the Supreme Court echoed the views of some proponents of evidence-based sentencing—notably that some level of prediction is already present in all aspects of the criminal justice system, such as bail decisions and parole hearings.[36] A similar case arose in *United States v. Fields*,[37] where the appellant did not argue that "psychiatric predictions of future dangerousness during the punishment phase are inadmissible *per se*" but challenged whether the expert opinion offered was reliably sufficient to be introduced at sentencing.[38] In the time between deciding *Barefoot* and *Fields*, the United States Supreme Court had ruled on *Daubert v. Merrell Dow Pharmaceuticals, Inc.*,[39] in which it interpreted Federal Rule of Evidence 702[40] and guided

---

30 *See* Brian Sites, Note, *The Danger of Future Dangerousness in Death Penalty Use*, 34 FLA. ST. U. L. REV. 959, 967–69 (2007).

31 *See* Eric S. Janus & Robert A. Prentky, *Forensic Use of Actuarial Risk Assessment with Sex Offenders: Accuracy, Admissibility and Accountability*, 40 AM. CRIM. L. REV. 1443, 1443–44 (2003).

32 *See infra* subpart II.A (Capital Cases) and subpart II.B (Sexually Violent Persons).

33 *See* Erica Beecher-Monas & Edgar Garcia-Rill, *Danger at the Edge of Chaos: Predicting Violent Behavior in a Post-*Daubert *World*, 24 CARDOZO L. REV. 1845, 1895–1901 (2003).

34 463 U.S. 880 (1983).

35 *See id.* at 899–904.

36 *See id.* at 897.

37 483 F.3d 313 (5th Cir. 2007).

38 *Id.* at 341.

39 509 U.S. 579 (1993). *Daubert* supplemented Federal Rule of Evidence 702, which had superseded *Frye v. United States*, 293 F. 1013 (D.C. Cir. 1923).

40 FED. R. EVID. 702.

courts acting as "gatekeepers" of expert testimony.[41] In *Fields*, the Fifth Circuit ruled that the Federal Death Penalty Act did not require a *Daubert* analysis, despite the American Psychological Association's urging that some sort of *Daubert*-like investigation should be carried out with regard to expert testimony in sentencing hearings.[42]

Even in the few states that have called for *some* gatekeeping test for expert future dangerousness predictions in death penalty cases, most have refused to categorically exclude future dangerousness predictions and many "exempt predictive expertise from any special testing, applying instead other rules on expert testimony and standard tests for relevance."[43] Only one judge has rejected predictive testimony under *Daubert* in a specially concurring opinion.[44] Most scholars agree that gatekeeping with regards to expert future dangerousness predictions in death penalty cases is necessary,[45] but disagree as to which standard to apply (that is, *Daubert* or a more relaxed standard) and whether future dangerousness predictions do, in

---

41 *See Daubert*, 509 U.S. at 600–601 (noting that even though "Rule 702 confides to the judge some gatekeeping responsibility in deciding questions of the admissibility of proffered expert testimony[,]" it does not "impose[ ] on them either the obligation or the authority to become amateur scientists in order to perform that role").

42 *Fields*, 438 F.3d at 342.

43 Alexander Scherr, Daubert & *Danger: The "Fit" of Expert Predictions in Civil Commitments*, 55 HASTINGS L.J. 1, 61–62, 73–75 (2003) (footnote omitted); *see* Sites, *supra* note 30, at 970–71 (noting that when courts are "faced with choosing between abandoning a desirable sentencing tool (future dangerousness) and relying on unsteady analytical tools (psychological testing), courts prefer the latter").

44 *See* Flores v. Johnson, 210 F.3d 456, 458–70 (5th Cir. 2000) (Garza, J., specially concurring) (subjecting the predictive testimony to a *Daubert* analysis and arguing that *Barefoot* should be overturned). *But see* People v. Murtishaw, 631 P.2d 446, 470 (Cal. 1981) (rejecting predictive dangerousness testimony not under *Kelly/Frye* but rather because the court found that the testimony's prejudicial impact outweighed its probative value). The court in *Murtishaw* noted that in places where the jury is required by law to determine whether someone is "dangerous," such predictive testimony, although unreliable, would often be the only evidence available to assist the trier of fact. *Id.* at 469; *see* 18 U.S.C. § 3593(c) (2012) ("Information is admissible regardless of its admissibility under the rules governing admission of evidence at criminal trials except that information may be excluded if its probative value is outweighed by the danger of creating unfair prejudice, confusing the issues, or misleading the jury.").

45 *See generally* Sites, *supra* note 30, at 987, 992–95 (arguing that *Daubert* should apply to death penalty sentencing proceedings and supplementary methods to minimize jury confusion); Erica Beecher-Monas, *The Epistemology of Prediction: Future Dangerousness Testimony and Intellectual Due Process*, 60 WASH. & LEE L. REV. 353, 364–79 (2003) (arguing that *Barefoot* was wrongly decided and proposing the constitutionalization of the *Daubert* test for death penalty sentencing proceedings).

fact, satisfy *Daubert*.[46] Despite the fact that *Daubert* does not apply to state or federal[47] sentencing procedures, academic and legal discussions have made it clear that *Daubert* offers a helpful framework through which to analyze the reliability of future dangerousness predictions in capital penalty sentencing.[48]

B. Sexually Violent Persons

Predictive assessments of future dangerousness have also been used in civil commitment hearings of sexually violent persons. Civil commitments are used to continue to incapacitate sex offenders who have finished their criminal sentences.[49] Risk assessment plays a central role in this assessment[50] and is often mandated by state law.[51]

Despite the known flaws of such predictive data techniques, courts have continued to allow this kind of evidence.[52] The legal and scholarly debate surrounding civil commitments highlights the lack of consensus over what standard of evidentiary admissibility should be used for expert testimony in civil commitment.[53] As in capital cases, some scholars have used *Daubert* as a framework to analyze the admissibility of future dangerousness predictions in sexual violent person civil commitment hearings.[54]

---

46 *See generally* Beecher-Monas & Garcia-Rill, *supra* note 33, at 1847, 1856–57, 1880, 1900 (applying *Daubert* as a constitutional ground to decide whether expert testimony on a specific defendant's future dangerousness violates due process and concluding that clinical predictions must be excluded and that actuarial instruments "barely squeak through" *Daubert*).

47 18 U.S.C. § 3661 (2012); U.S. SENTENCING GUIDELINES MANUAL § 6A1.3, Commentary (U.S. SENTENCING COMM'N 2016) ("In determining the relevant facts, sentencing judges are not restricted to information that would be admissible at trial.").

48 *See* Beecher-Monas & Garcia-Rill, *supra* note 33, at 1859.

49 *See* Janus & Prentky, *supra* note 31, at 1443.

50 *See id.* at 1443–44.

51 *See* George G. Woodworth & Joseph B. Kadane, *Expert Testimony Supporting Post-Sentence Civil Incarceration of Violent Sexual Offenders*, 3 LAW, PROBABILITY AND RISK 221, 236 (2004). For example, "Virginia's Sexually Violent Predator statute not only mandates the use of a specific instrument but also specifies the cutoff score on that instrument that must be achieved to proceed further in the commitment process." Jennifer L. Skeem & John Monahan, *Current Directions in Violence Risk Assessment*, 20 CURRENT DIRECTIONS IN PSYCHOL. SCI. 38, 38 (2011).

52 Janus & Prentky, *supra* note 31, at 1444.

53 *See* Woodworth & Kadane, *supra* note 51, at 236.

54 *See id.* (suggesting how risk assessments for violent sexual offenders may be changed to pass *Daubert* challenges); *see also* Janus & Prentky, *supra* note 31, at 1446 (arguing that actuarial risk assessments in civil commitments are admissible under *Daubert*).

C. Analogy to Evidence-Based Sentencing

The legal and academic discourse regarding future dangerousness predictions in capital cases and civil commitments highlights two competing goals—allowing the introduction of appropriate scientific evidence and giving the fact-finder as much information as possible to render an appropriate decision.[55] This tension also exists in evidence-based sentencing.

Despite the fact that sentencing hearings do not subject scientific evidence to the same rigorous testing as during trial,[56] this Note will follow the academic discussion surrounding future dangerousness predictions in capital cases and civil commitments and argue that evidence-based sentencing should pass *Daubert*.[57]

III
DAUBERT'S APPLICATION TO EVIDENCE-BASED SENTENCING GENERALLY AND IN PENNSYLVANIA

"*Daubert* unequivocally endorses 'empirically validated treatments' and 'evidence-based practices'" and protects the jury from giving undue weight to an unreliable and invalid expert testimony.[58] A *Daubert* analysis requires trial judges to decide "whether the reasoning or methodology underlying the

---

55 *See* Woodworth & Kadane, *supra* note 51, at 236 (proposing "that the statistical expert's job is to provide the fact-finder with the means to determine, as accurately and specifically as possible, the probability that this offender will recidivate given what is known"); *see also* Beecher-Monas & Garcia-Rill, *supra* note 33, at 1901 (noting that "[b]ecause juries are likely to evaluate the future dangerousness of any criminal defendant . . . they should be provided with the most accurate information that can bear on such an assessment. This includes actuarial risk factor studies . . . .").

56 One legal commentator characterizes the lack of safeguards around expert testimony at sentencing hearings as "an evidentiary free-for-all." Beecher-Monas, *supra* note 45, at 357. Nineteen states have ruled that either the states' evidentiary admissibility standards do not apply to expert testimony based on structured risk assessments or, if they do apply, these structured risk assessments meet the unchallenging standard. Daniel A. Krauss & Nicholas Scurich, *Risk Assessment in the Law: Legal Admissibility, Scientific Validity, and Some Disparities Between Research and Practice*, 31 BEHAV. SCI. L. 215, 220 (2013).

57 *But see generally* Pari McGarraugh, Note, *Up or Out: Why "Sufficiently Reliable" Statistical Risk Assessment Is Appropriate at Sentencing and Inappropriate at Parole*, 97 MINN. L. REV. 1079, 1083, 1112 (2013) (arguing for "sufficiently reliable" as an admissibility standard of actuarial risk assessments because of the other due process safeguards that are available to defendants at sentencing).

58 David L. Faigman & John Monahan, *Psychological Evidence at the Dawn of the Law's Scientific Age*, 56 ANN. REV. PSYCHOL. 631, 656 (2005); *see* Krauss & Scurich, *supra* note 56, at 219.

testimony is scientifically valid and . . . whether that reasoning or methodology properly can be applied to the facts in issue."[59]

Thus, testimony or evidence[60] must be reliable *and* relevant to help the trier of fact resolve the factual issue.[61] Legally,[62] reliability of the scientific technique turns on its scientific validity.[63] The relevance requirement under *Daubert* asks whether the information is helpful in deciding the case at hand.[64] This allows the trial court to analyze "whether this particular expert ha[s] sufficient specialized knowledge to assist the jurors 'in deciding the particular issues in the case.'"[65]

Despite acknowledging that there is no "definitive checklist or test,"[66] the United States Supreme Court laid out four factors typically relevant to determining reliability of scientific evidence: 1) "whether a theory or technique . . . can be (and has been) tested"; 2) "whether the theory or technique has been subjected to peer review and publication";[67] 3) what the known error rate (or potential rate of error) is and whether standards that control the technique's operation exist; and 4) whether the technique is generally accepted within a relevant scientific community and to what degree.[68] This analysis is a "flexible" inquiry focusing on "principles and methodology."[69]

---

59 Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 579, 592–93 (1993).

60 The *Daubert* test also includes expert testimony involving technical and other specialized knowledge. *See* Kumho Tire Co. v. Carmichael, 526 U.S. 137, 141 (1999) (holding "that *Daubert*'s general holding—setting forth the trial judge's general 'gatekeeping' obligation—applies not only to testimony based on 'scientific' knowledge, but also to testimony based on 'technical' and 'other specialized' knowledge").

61 *See Daubert*, 509 U.S. at 589 (stating that "under the Rules the trial judge must ensure that any and all scientific testimony or evidence admitted is not only relevant, but reliable"); Janus & Prentky, *supra* note 31, at 1460–61.

62 The author would like to draw attention to the fact that in the sciences, validity and reliability are two separate concepts. In such context, validity measures whether the answer given by a tool or test is correct. Reliability measures whether the answer given (regardless of correctness) is consistent. For a discussion of these differences, see DAVID B. ORR, FUNDAMENTALS OF APPLIED STATISTICS AND SURVEYS 54–55 (1995).

63 *See Daubert*, 509 U.S. at 590, n.9; Scherr, *supra* note 43, at 9.

64 *See* Scherr, *supra* note 43, at 9. For the sake of brevity, this Note will assume that the evidence-based sentencing is relevant, but readers are encouraged to consider Melissa Hamilton's article, *Adventures in Risk*, *infra* note 78, specifically her discussion on "Fit," as well as *infra* Part V (Public Policy Concerns).

65 *Kumho Tire Co.*, 526 U.S. at 156.

66 *Daubert*, 509 U.S. at 593.

67 *Id.* "[S]ubmission to the scrutiny of the scientific community is a component of 'good science,' in part because it increases the likelihood that substantive flaws in methodology will be detected." *Id.*

68 *Id.* at 594.

69 *Id.* at 594–95.

The following sections analyze evidence-based sentencing under each of the four *Daubert* factors for reliability. Each section draws on the literature surrounding actuarial assessments in general, followed by a closer analysis of the Pennsylvania Commission on Sentencing's proposed tool.

Although more than twenty states have embraced some use of evidence-based sentencing, they do so in very different ways.[70] In some jurisdictions, state legislatures have passed laws requiring actuarial risk assessments in sentencing procedures,[71] while in other jurisdictions, the judicial branch has endorsed risk assessments not required by the legislature.[72] To make matters more complicated, in still other jurisdictions, the judiciary and legislative branches have worked in tandem to create an evidence-based sentencing system.[73] After deciding to use evidence-based sentencing, states can either use commercial risk assessments, such as LSI-R and COMPAS, or create their own.[74] Therefore, the discussion below regarding evidence-based sentencing in general will be broad because a survey reviewing such diverse practices is outside of the scope of this Note.[75]

Finally, this Note argues that even if evidence-based sentencing were to pass a *Daubert* analysis, a mechanism such as Federal Rule of Evidence 403 would disallow the testimony. Federal Rule of Evidence 403 gives the Court "'very broad' discretion"[76] to exclude the expert evidence if the probative value

---

70 *See* Center for Sentencing Initiatives, *State Policies & Legislation: Interactive Map*, NCSC (July 2016), http://www.ncsc.org/microsites/csi/home/In-the-States/State-Activities/RNA-Map.aspx [https://perma.cc/3V7F-6XEK]; Luis Daniel, *The Dangers of Evidence-Based Sentencing*, GOVLAB (Oct. 31, 2014), http://thegovlab.org/the-dangers-of-evidence-based-sentencing/ [https://perma.cc/5X3G-AT6T].

71 Some such laws include TENN. CODE ANN. § 41-1-412 (2017); OHIO REV. CODE ANN. § 5120.114 (West 2011); OKLA. STAT. tit. 22 § 22-988.18 (West 2017); LA. STAT. ANN. § 15:327 (2014); COLO. REV. STAT. § 16-11-102 (2017); KY. REV. STAT. ANN. § 532.007 (West 2015); UTAH CODE ANN. § 77-18-1(3) (West 2017).

72 Some examples of these jurisdictions are Arkansas, Indiana, and Arizona. *See* Center for Sentencing Initiatives, *supra* note 70.

73 *Id.*

74 For example, Wyoming uses LSI-R, Michigan uses COMPAS, and Ohio has its own risk assessment system (ORAS). *Id.*

75 Though a comparison of different categories of current practices would significantly add to current literature, much of this information is inaccessible to the public.

76 United States v. Smithers, 212 F.3d 306, 322 (6th Cir. 2000) (quoting United States v. Hawkins, 969 F.2d 169, 174 (6th Cir. 1992)).

is substantially outweighed by a danger of unfair prejudice[77] or of confusing the jury.[78]

A. Introduction to the Pennsylvania Commission on Sentencing Risk Assessment Tool

The Pennsylvania Commission on Sentencing started developing a risk assessment tool in 2010.[79] The Commission developed and validated a risk assessment tool for all Level 3 and Level 4 offenses.[80] In 2015, the Commission published a report on the development of a new risk assessment scale for all five offense levels in Pennsylvania.[81] The results of this risk assessment tool (based on nine categories of Offense Gravity Scores for the risk of re-offense for any crime) was developed and validated twice using samples of offenders sentenced 1998-2000 and 2004-2006 and published in February 2016.[82] The Commission used the risk assessment tool created from 1998-2000 development sample, but not yet validated, to analyze the effects of removing sex, age, and county from the model.[83] Finally, in February 2016, the Pennsylvania Commission on Sentencing indicated its interest in creating a risk assessment tool to assess the risk of re-offense for a crime against a person.[84] Although the proposed law implementing the risk assessment instruments indicates that this tool has been created and will be used along with the risk assessment tool assessing the risk of re-offense for any crime,[85] the Commission

---

77 "Unfair prejudice 'does not mean the damage to a defendant's case that results from the legitimate probative force of the evidence; rather it refers to evidence which tends to suggest [a] decision on an improper basis.'" United States v. Bonds, 12 F.3d 540, 567 (6th Cir. 1993) (quoting United States v. Mendez-Ortiz, 810 F.2d 76, 79 (6th Cir. 1986)).

78 *See* Melissa Hamilton, *Adventures in Risk: Predicting Violent and Sexual Recidivism in Sentencing Law*, 47 ARIZ. ST. L.J. 1, 56–57 (2015) [hereinafter *Adventures in Risk*].

79 *See* PA. COMM'N ON SENT'G, RISK/NEEDS ASSESSMENT PROJECT INTERIM REPORT 1: REVIEW OF FACTORS USED IN RISK ASSESSMENT INSTRUMENTS 2 (2011).

80 *See* PA. COMM'N ON SENT'G, RISK ASSESSMENT PROJECT II INTERIM REPORT 2: VALIDATION OF A RISK ASSESSMENT INSTRUMENT BY OFFENSE GRAVITY SCORE FOR ALL OFFENDERS 2 (2016). [hereinafter VALIDATION BY OFFENSE GRAVITY SCORE].

81 *See* PA. COMM'N ON SENT'G, RISK/NEEDS ASSESSMENT PROJECT II INTERIM REPORT 1: DEVELOPMENT OF A RISK ASSESSMENT SCALE BY OFFENSE GRAVITY SCORE FOR ALL OFFENDERS 1 (2015).

82 *See id.* at 1–4.

83 *See* PA. COMM'N ON SENT'G, RISK ASSESSMENT PROJECT II SPECIAL REPORT: THE IMPACT OF REMOVING AGE, GENDER, AND COUNTY FROM THE RISK ASSESSMENT SCALE 2 (2015) [hereinafter IMPACT OF REMOVING DEMOGRAPHIC FACTORS].

84 *See* VALIDATION BY OFFENSE GRAVITY SCORE, *supra* note 80, at 28.

85 *See* PA. COMM'N ON SENT'G, PROPOSALS PUBLISHED IN PENNSYLVANIA BULLETIN (47 PA.B. 1999) § 305.2 (2017) [hereinafter PROPOSALS PUBLISHED], http://

has yet to release any of the data on the development or validation of such sample in its available reports.[86]

B. Prong 1: Whether Evidence-Based Sentencing Has Been Tested?

The first prong of the *Daubert* analysis requires determining whether evidence-based sentencing can be and has been tested.[87] This Note will first review the testing of common similar risk-assessment tools (though they are not necessarily only used for sentencing)[88] to show the difficulty with determining what is being tested and whether it can be tested. This Note will then turn to address the Pennsylvania Commission on Sentencing's risk assessment model.

Before deciding whether evidence-based sentencing can be and has been tested, it is necessary to ask what is being tested. This could include one of two questions: whether the statistical technique used to calculate a risk score has been tested, and whether the outcomes of such a statistical technique have been tested.[89]

As detailed in Part I, evidence-based sentencing relies on determining the predictive factors for a certain developmental

---

pcs.la.psu.edu/guidelines/proposed-for-public-comment-sentence-risk-assessment-instrument/annex-b/view [https://perma.cc/P5N9-5UV4].

86 The reports featured at the time this analysis was done are: Phase I Interim Report 1: Review of Factors Used in Risk Assessment Instruments; Phase I Interim Report 2: Recidivism Study: Initial Recidivism Information; Phase I Interim Report 3: Factors that Predict Recidivism for Various Types of Offenders; Phase I Interim Report 4: Development of Risk Assessment Scale; Phase I Interim Report 5: Developing Categories of Risk; Phase I Interim Report 6: Impact of Risk Assessment Tool for Low Risk Offenders; Phase I Interim Report 7: Validation of Risk Scale; Phase I Interim Report 8: Communicating Risk at Sentencing; Phase I Special Report: The Impact of Juvenile Record on Recidivism Risk; Phase II Interim Report 1: Development of a Risk Assessment Scale by Offense Gravity Score for All Offenders; Phase II Interim Report 2: Validation of Risk Assessment Instrument by OGS for All Offenses; Phase II Special Report: Impact of Removing Demographic Factors; Sentence Risk Assessment Instrument Adopted for Purpose of Public Comment. *See* PA. COMM'N ON SENT'G, *Overview of the Sentence Risk Assessment Instrument*, http://pcs.la.psu.edu/publications-and-research/research-and-evaluation-reports/risk-assessment [https://perma.cc/TY6S-36C8] (follow "Phase I," "Phase II," and "Sentence Risk Assessment Instrument for purposes of public comment" hyperlinks). Subsequent reports and drafts are still being added.

87 *See* Daubert v. Merrell Down Pharm., Inc., 509 U.S. 579, 593 (1993).

88 Many tests are wrongly used for sentencing despite specific warnings not to use them. *See, e.g.*, Starr, *supra* note 1, at 812 (citing DEP'T OF CORRECTIVE SERVS., LSI-R TRAINING MANUAL 8 (2002); BERNARD E. HARCOURT, AGAINST PREDICTION: PROFILING, POLICING, AND PUNISHING IN AN ACTUARIAL AGE 78–84 (2007); Starr, *supra* note 1, at 809 n.11).

89 *See* HANDBOOK OF PSYCHOLOGY IN LEGAL CONTEXTS 375–77 (David Carson & Ray Bull eds., 2d ed. 2003).

or normed sample, attributing appropriate weights for each predictive factor, and creating an estimated probability of the outcome (here, recidivism) occurring for each score or group of scores.[90] This is often achieved through a logistic regression, a commonly accepted statistical technique.[91] The Pennsylvania Commission on Sentencing used a logistic regression to create its actuarial instrument.[92]

Although using logistic regression to create predictive instruments is a commonly accepted and tested technique, a closer look at the Pennsylvania Commission on Sentencing's predictive instrument shows that not all logistic regressions are created equally. First, the predictive ability of a test is limited to the development sample.[93] For example, the Pennsylvania Commission on Sentencing randomly divided its samples of the offenders convicted between 1998 to 2000 and 2004 to 2006 at all five offense levels into two groups: one development and one validation.[94] The fact that the models derived from half of these two development groups were validated using the other half of the development groups in the same time frame does little to show its predictive value in the future, which is the purpose of this study. After creating and validating the model by two samples in the same time frame, the Commission should have gone one step further to test the predictive ability of the model on a sample that had not been used to develop the model, such as 2008 to 2010.

For the proposed law implementing these risk assessment tools, recidivism is measured by re-arrests within three years, which does not include out-of-state, federal or foreign charges, or a re-arrest *dismissed* by a minor court.[95] Although the Pennsylvania Sentencing Commission included technical viola-

---

90 *See Adventures in Risk*, *supra* note 78, at 9. Within the group of score known as "risk bins" (high, medium, low), there is no commonly agreed upon definitions of risk categories, accepted metrics, or normative legal distinctions for such labels. Among different tests, the same person can have a vastly different risk category. *See, e.g.*, *id.* at 21 (describing a study in which sex offenders received disparate categorical labels based on scores from different risk assessment instruments).

91 Personal communication with Professor John Zipp, University of Akron (Jan. 2, 2017). *See* GARETH JAMES ET AL., AN INTRODUCTION TO STATISTICAL LEARNING: WITH APPLICATIONS IN R 127 (2013, corrected at 8th printing 2017).

92 *See* PA. COMM'N ON SENT'G, INTERIM REPORT 3: FACTORS THAT PREDICT RECIDIVISM FOR VARIOUS TYPES OF OFFENDERS (PHASE I) 6 (2011) [hereinafter FACTORS THAT PREDICT RECIDIVISM].

93 *See* HANDBOOK OF PSYCHOLOGY IN LEGAL CONTEXTS, *supra* note 89, at 381.

94 *See* VALIDATION BY OFFENSE GRAVITY SCORE, *supra* note 80, at 9.

95 PROPOSALS PUBLISHED, *supra* note 85, § 305.1(b)(15).

tions as re-arrests in its development sample,[96] the final proposed law explicitly excludes the use of the Sentence Risk Assessment Instrument for the sentences imposed as a result of "revocation of probation, intermediate punishment or parole."[97] Depending on the quantity and shared characteristics of those re-arrested on a technical violation, the current risk assessment tool is likely to fit less well due to this statutory exclusion that was not accounted for in the creation of the model.

Additionally, arrest itself may be a false proxy for criminality resulting from higher level of policing in certain neighborhoods.[98] The high incidence of plea bargaining in the U.S. criminal justice system, which may include a significant amount of innocent people who plead guilty even though they could have had their case dismissed by a minor court, compounds concerns that arrest is an inappropriate measurement of recidivism.[99]

In the end, the results from an actuarial risk assessment (usually a number on a scale) do little to determine the qualitative features of the defendant's punishment (jail, probation, or fine; length of time; and possibility of parole). The Commission attempts to contextualize the recidivism risk by using two separate evidence-based sentencing tools for "Risk of Re-Offense for Any Offense" and the "Risk of Re-Offense for an Offense

---

96 *See* VALIDATION BY OFFENSE GRAVITY SCORE, *supra* note 80, at 4.

97 PROPOSALS PUBLISHED, *supra* note 85, § 305.3(b).

98 When "there are more police officers making more arrests in high-crime neighborhoods," those living in that area "will automatically score higher simply because of where they live." Therefore, when police are predisposed to arrest minorities without good cause, a prediction based on arrests can falsely interpret arbitrary arrests as a sign of increased criminality among minority groups, which creates a positive feedback loop that reinforces the pattern. *See* Ryan Briggs, *Should Your Race or ZIP Code Determine Jail Time?*, CITY & STATE PA. (Aug. 2, 2016, 11:39 AM), http://www.cityandstatepa.com/content/brewing-battle-over-what-factors-will-determine-jail-time [https://perma.cc/AMW7-LTZC]. *See also* Testimony of the Defender Association of Philadelphia, Pennsylvania Association of Criminal Defense Lawyers, Public Defender Association of Pennsylvania and the Allegheny County Public Defender's Office Before the Sentencing Commission of Pennsylvania (May 25, 2017), at 3–4, http://pcs.la.psu.edu/guidelines/sentencing/sentencing-guidelines-and-implementation-manuals/7th-edition-amendment-4-1-1-2018-1/7th-edition-amendment-4-adopted-for-purposes-of-public-comment/testimony/testimony-mark-houldin-and-bradley-s.-bridge-defender-association-of-philadelphia-philadelphia-may-25-2017/view [https://perma.cc/G7EW-72RB] (describing the racial disparity in policing measures in Pennsylvania).

99 *See* Jed S. Rakoff, *Why Innocent People Plead Guilty*, N.Y. REV. OF BOOKS (Nov. 20, 2014), http://www.nybooks.com/articles/2014/11/20/why-innocent-people-plead-guilty/?insrc=whc [https://perma.cc/S6KS-GGKB].

Against a Person";[100] however, the contours and relative gravity of these crimes are hazy. A robbery of a motor vehicle is considered an "Offense Against a Person,"[101] but operating a meth lab is not.[102] Similarly, gravity between and among these types of crimes is not intuitive; for example, both fortunetelling for money and operating a meth lab are considered "Any Offense."[103] Thus, the Pennsylvania Commission on Sentencing's predictions tell us the probability a person will be rearrested within the next three years for a wide range of crimes based on a tool that has not been validated on a future sample but instead only on a sample in the same time frame as the model's development sample.

Although choosing the right sample upon which to create a predictive model through logistic regression is crucial, the Pennsylvania Commission on Sentencing's methods also demonstrate how the selection of the model's variables is important. When selecting the factors to include in its model, Pennsylvania's risk assessment tool only took the factors found to be significant predictors of recidivism.[104] Although race and county were statistically significant factors,[105] the final Pennsylvania Commission on Sentencing risk assessment tool does not include race and county as factors.[106] The Commission explains that race and county are "statistically controlled for in the analyses, which means that the effects of the other factors are included only after eliminating the effects of race and county."[107] The Commission never explains why it keeps race and county in the model to create the risk assessment tool even when it knows that it is going to exclude it from the risk scale. Thus, it is preferable that the logistic regression model be rerun without race and county rather than controlling for it post hoc. By including both race and county in the model, the Commission could be ignoring other factors that could have been

---

100 *See* PROPOSALS PUBLISHED, *supra* note 85, § 305.2(b)(2).

101 *See* PROPOSALS PUBLISHED, *supra* note 85, § 305.1(b)(9) (citing 42 PA. CONS. STAT. § 9714(g) (2015)).

102 Given that the definition of § 305.1(b)(9) in PROPOSALS PUBLISHED, *supra* note 85, is exclusive, operating a meth lab would not be considered an "Offense Against a Person" as it does not fall within any of the three categories of crimes defining an "Offense Against a Person." *See* 18 PA. CONS. STAT. § 7508.2 (2015) (operating a meth lab).

103 Because neither is included in the definition of "Offense Against a Person," they both fall into the residual category of "Any Offense." *See* 18 PA. CONS. STAT. § 7104 (2015) (fortunetelling).

104 *See* VALIDATION BY OFFENSE GRAVITY SCORE, *supra* note 80, at 12.

105 *See id.* at 51–58 ("Final Logistic Models by Offense Gravity Score").

106 *See* PROPOSALS PUBLISHED, *supra* note 85, § 305.1(b)(18).

107 VALIDATION BY OFFENSE GRAVITY SCORE, *supra* note 80, at 12 n.8.

significant in the development model had race and county not been initially included.[108]

The outcomes of such predictive instruments have been typically measured using area under the curve (AUC) analysis to measure the discrimination of the test; that is, how well the predictive test can differentiate those who experienced the outcome of interest (labeled as likely to recidivate and did recidivate) versus those who did not (labeled as likely to recidivate but did not recidivate).[109] Melissa Hamilton notes that many researchers misconstrue the meaning of the area under the curve when evaluating risk assessment tools. The AUC analysis provides "'the probability that a randomly selected individual who committed an [act of recidivism] . . . received a higher risk classification than a randomly selected individual who did not' reoffend."[110] This does not tell us "information on the accuracy of any individual prediction,"[111] "the probability that individuals are scored correctly . . . [, or] the potential that a person assessed with a high test score will eventually become a recidivist."[112]

Typically, AUC values in popular risk assessments for violent and sexual recidivism tools range from 70-75%, meaning that "these risk instruments have been able to classify violent and sexual recidivists at higher levels of risk than non-recidivists about 70 to 75% of the time."[113] Although these values

---

108 Personal Communication with Professor John Zipp (Jan. 2, 2017). *See also generally Adventures in Risk*, *supra* note 78, at 20–21 (discussing how "risk tools typically include a relatively small number of variables, thereby omitting a plethora of potential explanatory or correlative factors").

109 *See Adventures in Risk*, *supra* note 78, at 25.

110 *Id.* at 25 (alteration in the original) (quoting Jay P. Singh et al., *Measurement of Predictive Validity in Violence Risk Assessment Studies: A Second-Order Systematic Review*, 31 BEHAV. SCI. & L. 55, 64 (2013)). "An AUC of .90, as an illustration, means that if one randomly chooses a recidivist and a non-recidivist, the recidivist's actuarial score would be higher than the non-recidivist's score about 90% of the time." *Id.* (citing Christopher T. Lowenkamp et al., *The Federal Post Conviction Risk Assessment (PCRA): A Construction and Validation Study*, 10 PSYCHOL. SERVS. 87, 92 n.11 (2013)). More problematic is that "a scale can achieve a high rating for discrimination even when the average predicted risk of violent reoffense is significantly different than the actual percentage of violent recidivists." *Id.* at 24 (citing Nancy R. Cook, *Use and Misuse of the Receiver Operating Characteristic Curve in Risk Prediction*, 115 CIRCULATION 928, 928 (2007)).

111 *Id.* at 27 (citing Thomas Nilsson et al., *The Precarious Practice of Forensic Psychiatric Risk Assessments*, 32 INT'L J.L. & PSYCHIATRY 400, 402 (2009)).

112 *Id.* at 26 (citing Cook, *supra* note 110, at 928).

113 *Id.* Some experts contend that this common rate is due to similar factors being used by recidivism risk tools and that technologies have possibly reached the natural limit for predicting human behavior. *Id.* at 28 (citing John Monahan & Jennifer L. Skeem, *Risk Redux: The Resurgence of Risk Assessment in Criminal Sanctioning*, 26 FED. SENT'G REP. 158, 158 (2014)).

perform statistically better than chance (50%), many researchers equate these values to moderate or large effect sizes.[114] However, in statistics, there is no consensus as to which AUC scores represent small, moderate, or large effect sizes.[115] Without such a consensus, researchers could tout that a certain predictive instrument effect size is large enough to be used in the courts without having to define what "large" is across the literature or describe the specific definition of a tool's AUC.[116]

Finally, the Commission confines the use of this tool in a way that may not fully achieve the benefits of evidence-based sentencing, particularly reducing incarceration rates. For offenders falling into the typical risk category (the middle 68%), the Commission makes no additional recommendation; therefore these defendants are sentenced on the information available.[117] For low- or high-risk offenders, the Commission recommends a separate pre-sentencing report to collect more information before sentencing.[118] Whether more closely analyzing the remaining 32% of low- or high-risk offenders' situations will actually result in different, more effective incarceration is an open question. Although using logistic regressions to predict individual outcomes as a general technique can be and has been tested, the Pennsylvania Commission on Sentencing fails to justify why its sample, validation method, and variables differ from what one would expect in a similar scenario.

C. Prong 2: Whether Evidence-Based Sentencing Has Been Subjected to Peer Review and Publication?

Judges typically pay close attention to publication and peer review, even more so than the testing or error rates.[119] Judges may focus on journal names as a proxy for quality scholarship, but it is important to assess the content of the published studies.[120] The importance of peer review is to de-

---

114 *Id.* at 26 (citing R. Karl Hanson & David Thornton, *Improving Risk Assessments for Sex Offenders: A Comparison of Three Actuarial Scales*, 24 LAW & HUM. BEHAV. 119, 129 (2000)).

115 *Id.*

116 Hamilton suggests that focusing on calibration is a better benchmark for evaluating an instrument's predictive ability. *Id.* at 28–29.

117 *See* PROPOSALS PUBLISHED, *supra* note 85, §§ 305.1(b)(17), 305.5(b).

118 *See* PROPOSALS PUBLISHED, *supra* note 85, § 305.5(c).

119 Jay P. Kesan, *An Autopsy of Scientific Evidence in a Post-*Daubert *World*, 84 GEO. L.J. 1985, 2029 (1996).

120 *Id.* at 2029; *Adventures in Risk*, *supra* note 78, at 57–61. *See generally* MICHAEL J. SAKS ET AL., ANNOTATED REFERENCE MANUAL ON SCIENTIFIC EVIDENCE

tect the design flaws of studies; any conclusion is only as good as the methods used to attain it.[121]

Generally, evidence-based sentencing has received policy-based and learning-oriented discussions.[122] Peer reviews specific to the technique and efficacy of actuarial risk assessments, in general, are commonly limited to predictions of violence,[123] not recidivism. With its growing use, however, there is an increasing presence of academic literature surrounding risk assessment in sentencing.[124]

Several factors impede the ability of the scientific community to peer review methods specific to validity. First, the proprietary nature of commercial tests creates conditions in which the company producing the test is validating its own studies or is involved in another researcher's validation.[125] This self-review does not count as peer review, and those reviewing it on behalf of the company should disclose this potential conflict of interest. Additionally, limited data from states where evidence-based sentencing has been used for a long time, such as Virginia, make it difficult to evaluate both the methods and efficacy.[126]

Not only does the proprietary nature of commercial tools impede the ability of scientists to conduct peer review on the

---

107–08 (2d ed. 2004) (discussing several scenarios where a position's illusion of prestige masks flaws in their methods).

121 *See* Kesan, *supra* note 119, at 2029–30.

122 *See, e.g.*, PAMELA M. CASEY ET AL., NAT'L CTR. FOR STATE COURTS, OFFENDER RISK & NEEDS ASSESSMENT INSTRUMENTS: A PRIMER FOR COURTS 2 (2014), [hereinafter OFFENDER RISK & NEEDS ASSESSMENT INSTRUMENTS] http://www.ncsc.org/~/media/Microsites/Files/CSI/BJA%20RNA%20Final%20Report_Combined%20Files%208-22-14.ashx [https://perma.cc/MXU7-X5NZ] ("This Primer is a resource to help judges and others involved in sentencing understand and make knowledgeable decisions about the value and use of a [risk and needs] assessment.").

123 *See* Jay P. Singh et al., *Measurement of Predictive Validity in Violence Risk Assessment Studies: A Second-Order Systematic Review*, 31 BEHAV. SCI. L. 55, 55 (2013) ("[T]hese [risk assessment] instruments are designed to aid in the assessment of risk for antisocial behavior, most commonly general violence, sexual violence, and criminal offending."); Min Yang et al., *The Efficacy of Violence Prediction: A Meta-Analytic Comparison of Nine Risk Assessment Tools*, 136 PSYCHOL. L. BULL. 740, 741 (2010) ("In most cases, actuarial tools are designed by combining empirically or theoretically derived constructs that are predictive of violence or antisocial activities to guide the forecasting of future antisocial or violent acts.").

124 Skimming through the following sources helps give an idea of the amount of publications. Letter from Brian D. Johnson, Professor at Univ. of Md., James P. Lynch, Professor at Univ. of Md., & Rebecca Richardson, Doctoral Student, to Conn. Sentencing Comm'n (Dec. 1, 2015), http://www.ct.gov/ctsc/lib/ctsc/2016-01DRAFT.pdf [https://perma.cc/52RG-TP8P]; OFFENDER RISK & NEEDS ASSESSMENT INSTRUMENTS, *supra* note 122, at 32.

125 *See* Starr, *supra* note 8 (noting that risk prediction instruments are frequently not very transparent).

126 *The New Science of Sentencing*, *supra* note 24.

tool, it also interferes with due process considerations. In March 2017, the United States Supreme Court indicated that it was interested in whether a defendant's right to due process was violated by a judge's consideration of a COMPAS report that the defendant was unable to inspect or challenge by requesting that the federal government file an *amicus* brief arguing whether it should hear the case.[127]

Although there is a growing body of peer review literature for evidence-based sentencing, Hamilton's critique of incorrect statistical interpretation warns of the danger of reading too quickly.[128] Under *Daubert*, courts should consider which test is being presented when looking for specific peer review, as not all research on risk assessment will be applicable. Although the Pennsylvania Commission on Sentencing has said that it will seek an external validation once the development process is completed, if it has occurred, it is not yet available, and the risk assessment program is set to begin on July 1, 2018 without any restrictions subjecting the risk assessment tools to external peer review.[129]

D. Prong 3: What Is the Known (or Potential) Error Rate and Are There Standards that Control Evidence-Based Sentencing?

The error rates of evidence-based sentencing are crucial to the *Daubert* analysis and directly hinge on how much error a society is willing to tolerate in exchange for the purported safety benefits of correctly identifying a defendant with a high-risk of recidivism.[130] Evidence-based sentencing, by churning

---

127 Adam Liptak, *Sent to Prison by a Software Program's Secret Algorithms*, N.Y. TIMES (May 1, 2017), https://www.nytimes.com/2017/05/01/us/politics/sent-to-prison-by-a-software-programs-secret-algorithms.html [perma.cc/VR2Z-J422].

128 *Adventures in Risk*, *supra* note 78, at 2–3.

129 The fact that Pennsylvania is seeking an external peer review is available through a source created by the former Chair of the Pennsylvania Commission on Sentencing. No other information on the status of this external validation was posted on the Commission on Sentencing's risk assessment website. Steven L. Chanenson & Jordan M. Hyatt, *The Use of Risk Assessment at Sentencing: Implications for Research and Policy* 9 (Villanova U. Charles Widger Sch. of L. Working Paper Series, 2016) http://digitalcommons.law.villanova.edu/cgi/viewcontent.cgi?article=1201&context=wps [https://perma.cc/G855-3ZDP].

130 Some try to draw false negative parallels with a comparison between medicine and law, but law has different standards and goals than medicine. *See Adventures in Risk*, *supra* note 78, at 55. *But see,* Aaron E. Carroll, *What We Mean When We Say Evidence-Based Medicine*, N.Y. TIMES (Dec. 27, 2017), https://www.nytimes.com/2017/12/27/upshot/what-we-mean-when-we-say-evidence-based-medicine.html [https://perma.cc/988D-KHK6] (explaining how even "evi-

out a single number on a scale interpreted as an indication of a defendant's future recidivism rate, has the potential to mask the risk of error with the allure of the scientific method.[131]

Each risk assessment tool had its own calculable error rate, which means that each tool should be individually subject to *Daubert*. However, to provide some context to the general efficacy of risk assessments, a meta-analysis found that nine different violence risk assessment tools had comparable, moderate levels of predictive efficacy and recommended that they not be used as the primary means for a criminal justice decision.[132]

Another problem that aggravates errors is the failure to conduct local validations of commercial methods in approximately 70 percent of the jurisdictions that use assessment instruments, which can result in inaccurate classifications.[133] Local validations are important because of the difference between jurisdictions as to their "laws, policies, or sentencing guidelines (e.g., management of felony offenses, eligibility for probation, risk of recidivism defined locally as re-arrest vs. reconviction) may create a unique set of circumstances and constraints that can reduce the effectiveness of a tool created elsewhere."[134]

For Pennsylvania, the error rate depending upon the distribution of the risk categories led to an overall prediction accuracy ranging from 49 to 65 percent for the tools developed for the Level 3 and Level 4 samples.[135]

> The lower the cutoff point for determining high risk, the better the prediction accuracy for high risk . . . but the worse the prediction accuracy for low risk . . . . Alternatively, the higher the cutoff point for determining high risk, the worse the prediction accuracy for high risk, but the better the prediction accuracy for low risk.[136]

---

dence-based medicine" is facing similar definitional, practical, and policy concerns).

131 *See id.* at 56–57.

132 *See* Yang et al., *supra* note 123, at 761.

133 PAMELA M. CASEY ET AL., NAT'L CTR. FOR STATE COURTS, USING OFFENDER RISK AND NEEDS ASSESSMENT INFORMATION AT SENTENCING: GUIDANCE FOR COURTS FROM A NATIONAL WORKING GROUP 29–31 (2011), http://www.ncsc.org/~/media/Microsites/Files/CSI/RNA%20Guide%20Final.ashx [perma.cc/XU2U-TXEB].

134 *Id.* at 30 (emphasis omitted).

135 PA. COMM'N ON SENT'G, RISK/NEEDS ASSESSMENT PROJECT INTERIM REPORT 7: VALIDATION OF RISK SCALE 2, 4 (2013).

136 *Id.* at 4.

The Commission does not provide a similarly user-friendly analysis for its current risk assessment tool that includes all of the levels of offense.[137]

E. Prong 4: Whether Evidence-Based Sentencing Is Generally Accepted Within a Relevant Scientific Community and to What Degree?

The purpose of the general acceptance prong of *Daubert* is to uncover whether there is general agreement that the scientific theory is "not based on a novel theory or procedure that is 'mere speculation or conjecture.'"[138] A theory that does not have the acceptance of most of the pertinent scientific community or is disfavored by a substantial part of the scientific community will not be generally accepted.[139] Although absolute certainty is not necessary for a finding of general acceptance,[140] it is crucial to first determine the scope of the relevant scientific community in order to establish whether evidence-based sentencing is generally accepted.[141] The criteria used to decide who is part of the relevant scientific community must be broad enough to reach a spectrum of scientists who are not entirely financially, professionally, and ideologically invested in the field.[142] However, in the search to find a broad enough scientific community, the factfinder must still limit the community to those who truly understand the research basis for the work.[143] Though judges, parole officers, pre-sentence investigators, corrections officers, and other actors in the criminal justice system are exposed to and use a variety of actuarial risk

---

137 VALIDATION BY OFFENSE GRAVITY SCORE, *supra* note 80, at 22.

138 United States v. Bonds, 12 F.3d 540, 562 (6th Cir. 1993) (quoting United States v. Brown, 557 F.2d 541, 559 (6th Cir. 1977)).

139 *Id.* at 562 (citing Novak v. United States, 865 F.2d 718, 725 (6th Cir. 1989)).

140 *Id.*

141 David L. Faigman et al., *Group to Individual (G2i) Inference in Scientific Expert Testimony*, 81 U. CHI. L. REV. 417, 460, 463 (2014) [hereinafter *G2i*] ("The key difficulty [ ] involves identifying what field should be selected for review. Very often fields are defined by self-interested practitioners or established guilds. For example, if a court asks experts in the areas of polygraphs, bite marks, bullet lead, or hair identification about general acceptance, it would likely hear a chorus of consensus, though each of these areas of claimed expertise has been thoroughly discredited.").

142 *Id.* at 461 ("Professional overinvestment might include a researcher who is a leading figure in the field but whose life's work depends on acceptance of the expertise. Ideological investment might include a researcher whose judgment about the validity of an empirical framework will be influenced by its ability to further a desired policy outcome.").

143 *Id.* at 461–62.

assessments in making decisions, they are not the relevant scientific community.[144]

For evidence-based sentencing, the relevant scientific community could extend to statisticians and social scientists, but should exclude, or limit the weight given to, statisticians or researchers who are engaged for financial reasons or those who engage with actuarial risk assessments for purely policy reasons. The first limitation on financial motivations should discount evidence from researchers who have created actuarial risk assessments tools for commercial uses as well as those who have tested these tools using funds from a commercial provider. The second limitation of ideological investment may be more difficult to outline, given that many researchers creating specific tools are doing so with the hopes that it will reduce incarceration rates.

Courts have typically noted that general acceptance is needed with respect to both the theory and the methodology of the technique.[145] Evidence-based sentencing uses group statistics to create an individual prediction. This group to individual prediction (G2i) is common throughout many different fields.[146] The methodology of evidence-based sentencing involves the development of the sample to create the predictive test, the variables chosen, and the interpretation of results typically through logistic regression,[147] which, given the amount of diverse predictive tools,[148] has been generally accepted. At its broadest, logistic regression resulting in individual prediction of an individual is generally accepted as an appropriate technique when analyzing a dichotomous dependent variable.[149]

---

144 *See* Williamson v. Reynolds, 904 F. Supp. 1529, 1558 (E.D. Okla. 1995) (noting that general acceptance of hair analysis should not be based off "hair experts who are generally technicians testifying for the prosecution," but rather "scientists who can objectively evaluate such evidence").

145 United States v. Bonds, 12 F.3d 540, 562 (6th Cir. 1993).

146 *See G2i*, *supra* note 141, at 421–22 (describing G2i examples such as in medicine, where research regarding at what age women should begin having annual mammograms, provides an empirical framework to help doctors make individual decisions and help manage the risks of breast cancer in individual patients, and in meteorology, where group data models the trajectory and severity of storms to help determine whether to evacuate a community due to the threat of a particular storm).

147 *See, e.g.*, Chanenson & Hyatt, *supra* note 129, at 9 (explaining the methodology of Pennsylvania's evidence-based sentencing).

148 PA. COMM'N ON SENT'G, RISK/NEEDS ASSESSMENT PROJECT INTERIM REPORT 1: REVIEW OF FACTORS USED IN RISK ASSESSMENT INSTRUMENTS app. C (2011).

149 Personal communication with John Zipp (Jan. 2, 2017). *See* JEFFERY T. ULMER, SOCIAL WORLDS OF SENTENCING: COURT COMMUNITIES UNDER SENTENCING GUIDELINES 43 (Austin T. Turk ed., 1997) ("Although other techniques for modeling

Though not a legal question, what remains to be generally accepted is the question: Who should decide the *policy* rationales for evidence-based sentencing, and which goals of evidence-based sentencing to promote?

IV
DAUBERT'S INTERACTION WITH FEDERAL RULE OF EVIDENCE 403

Interestingly, most courts tend not to consider the *Daubert* factors in a step-by-step analysis.[150] Instead, one study found that the best predictor of whether evidence would be admissible was, *inter alia*, whether the evidence was prejudicial. This section focuses on prejudicial information through the lens of Federal Rule of Evidence 403.

Federal Rule of Evidence 403 acts as a final check on the *Daubert* analysis. Even if a scientific technique passes the *Daubert* analysis of reliability and fit, Federal Rule of Evidence 403 may disallow testimony if the probative value is substantially outweighed for many reasons.[151] With regards to evidence-based sentencing, the danger of creating unfair prejudice or confusing the jury would substantially outweigh its probative value. Evidence-based sentencing creates a risk of unfair prejudice by using immutable factors outside of the defendant's control.[152] It also confuses the jury members, who do not engage in appropriate critical inquiries about its scientific validity.[153]

Although a typical evidence-based sentencing analysis contains many questionable factors unrelated to the crime,[154] the subsequent analysis will focus on sex and age. These two categories are analyzed for two reasons. First, the Pennsylvania Commission on Sentencing analyzed the efficacy of its

---

dichotomous dependent variables exist . . . logistic regression is widely held to be the most appropriate procedure.").

150 A. Leah Vickers, Daubert, *Critique and Interpretation: What Empirical Studies Tell Us About the Application of* Daubert, 40 U.S.F. L. REV. 109, 133 (2005).

151 FED. R. EVID. 403 ("The court may exclude relevant evidence if its probative value is substantially outweighed by a danger of one or more of the following: unfair prejudice, confusing the issues, misleading the jury, undue delay, wasting time, or needlessly presenting cumulative evidence.").

152 *See* IMPACT OF REMOVING DEMOGRAPHIC FACTORS, *supra* note 83, at 1. *See also* Starr, *supra* note 1, at 829 ("[A] generalization about a behavioral tendency like criminal recidivism is simply not comparable to a physical difference . . . [and] certain kinds of generalizations (including those concerning gender) are particularly socially harmful or expressively invidious, even if they have statistical support.").

153 *Adventures in Risk*, *supra* note 78, at 56–57.

154 Starr, *supra* note 1, at 805.

evidence-based sentencing model when demographic factors were removed, thus showing the relative "costs" in terms of accuracy of not including such questionable factors.[155] Because the Commission only conducted this analysis on the risk assessment tool based on the 1998 to 2000 development model (and not the tool used for the final proposed risk assessment),[156] the figures discussed below may not correspond exactly to the current proposed risk assessment tools. Regardless, the general concepts, critiques, and concerns remain the same.

Additionally, each of these categories touches on a constitutional issue or an issue of concern for the court. The inclusion of sex in evidence-based sentencing creates an equal protection issue subject to heightened scrutiny.[157] Age, although not a protected class and subject to only rational basis review, has been considered by the Supreme Court, particularly when focusing on the sentencing of juvenile offenders.[158]

It should be noted that the Pennsylvania Commission on Sentencing's review of these factors was prompted by the constitutional, ethical, and fairness issues raised by Sonja Starr in *Sentencing by Numbers* and *Evidence-Based Sentencing and the Scientific Rationalization of Discrimination*.[159] Despite the concerns and their moderate accuracy, the Pennsylvania Commission on Sentencing has recommended keeping these factors in current and future risk assessment models.[160]

A. Sex

Sex is commonly used in evidence-based sentencing and results in harsher sentences for men.[161] Within the Pennsylvania Commission on Sentencing's eight Offender Gravity

---

155 IMPACT OF REMOVING DEMOGRAPHIC FACTORS, *supra* note 83, at 1.

156 VALIDATION BY OFFENSE GRAVITY SCORE, *supra* note 80, at 21–22.

157 Craig v. Boren, 429 U.S. 190, 197 (1976) ("[C]lassifications by gender must serve important governmental objectives and must be substantially related to achievement of those objectives.").

158 *See* Miller v. Alabama, 567 U.S. 460, 489 (2012) (holding that mandatory life without parole for a juvenile convicted of murder is unconstitutional); Graham v. Florida, 560 U.S. 48, 79–80 (2010) (ruling that life imprisonment without parole for non-homicide offenses committed by juveniles is unconstitutional); Roper v. Simmons, 543 U.S. 551, 570–71 (2005) (holding that the death penalty for juveniles is unconstitutional due to their lack of culpability relative to adults).

159 IMPACT OF REMOVING DEMOGRAPHIC FACTORS, *supra* note 83, at 1.

160 *Id.* at 1–2.

161 *See* Massie, *supra* note 25, at 522.

Categories[162] that use sex in their risk assessment, sex accounts for 1-5% of the risk model's prediction of recidivism.[163] In evaluating whether sex could be removed from the risk assessment model, the Pennsylvania Commission on Sentencing concluded, "[E]ven though removing [sex] . . . would impact the utility of the scale to a much lesser extent [than age],[164] since [it] do[es] provide some increase in the recidivism prediction, it would be worth keeping [it] in the scale as well."[165] Removing sex would result in females appearing to have higher recidivism rates, and males appearing to have lower recidivism rates, as opposed to how these statistics appear in the original scale.

In its most recent validation of Risk Assessment Instrument by Offense Gravity Score [OGS] for All Offenders for the 2004-2006 development sample, male defendants all scored one point higher than female defendants.[166] Given that the total risk scores range from eight to nineteen depending upon the defendant's OGS category, sex carries a substantial weight on the severity of one's sentence.

Despite the fact that the Federal Sentencing Guidelines explicitly forbid sex as a sentencing consideration and that case law regarding the use of sex in non-evidence-based sentencing proceedings is clear,[167] the revised Model Penal Code encourages the inclusion of sex as a factor while offering no equal protection analysis as to why it would be acceptable.[168]

---

162 The current proposed risk assessment tool for "Any Crime" and "Crime Against a Person" take sex into consideration for seven and six Offender Gravity Categories, respectively. PROPOSALS PUBLISHED, *supra* note 85, § 305.7(a)-(b).

163 IMPACT OF REMOVING DEMOGRAPHIC FACTORS, *supra* note 83, at 5.

164 *See infra* subpart IV.B discussion on Age.

165 IMPACT OF REMOVING DEMOGRAPHIC FACTORS, *supra* note 83, at 1.

166 VALIDATION BY OFFENSE GRAVITY SCORE, *supra* note 80, at 19.

167 U.S. SENTENCING GUIDELINES MANUAL § 5H1.10 (U.S. SENTENCING COMM'N 2012). The few times this issue has arisen, courts have ruled that basing sentences on sex is unconstitutional. *See* Starr, *supra* note 1, at 824 (citing United States v. Maples, 501 F.2d 985, 987 (4th Cir. 1974); Williams v. Currie, 103 F. Supp. 2d 858, 868 (M.D.N.C. 2000); Carissa Byrne Hessick, *Race and Gender as Explicit Sentencing Factors*, 14 J. GENDER RACE & JUST. 127, 137, n.68 (2010)).

168 MODEL PENAL CODE: SENTENCING § 6B.09 reporter's note at 62 (AM. LAW INST., Tentative Draft No. 2, 2011); Starr, *supra* note 1, at 824 (noting that "the MPC drafters . . . offer no commentary defending [the inclusion of sex] on constitutional grounds, as though its constitutionality is self-evident"). The Model Penal Code does note that referring to one's race or ethnicity during sentencing would be unconstitutional. Despite the omission of a seemingly important distinction between the difference in constitutionality of race and sex, the Code was passed in 2011 and remains the same in the Third Tentative Draft. MODEL PENAL CODE: SENTENCING xii (AM. LAW INST., Tentative Draft No. 3, 2014).

This is troubling, because an equal protection analysis would suggest otherwise.[169] Under intermediate scrutiny, distinctions made on the basis of sex must serve "important governmental objectives and [] the discriminatory means employed [must be] substantially related to the achievement of those objectives."[170] This discrimination cannot be based on "overbroad generalizations" about the differences between men and women.[171] Even when data support an even stronger predictive empirical relationship than evidence-based sentencing, the Supreme Court has typically found the discriminatory means to be unconstitutional.[172] This consistent rejection of a sex-based distinction calls into question its use during the sentencing procedure.

B. Age

Age is the most important demographic factor predicting arrest.[173] In fact, the Pennsylvania Commission on Sentencing found that for each year of increase in age, the likelihood of

---

169 *See* Testimony of the Defender Association of Philadelphia, *supra* note 98, at 5 (focusing on the Pennsylvania Risk Assessment Instrument and arguing that it is "unconstitutional to attach demerit points simply because of one's gender"). Some scholars argue that considering a variable subject to intermediate or heightened scrutiny along with other variables could be acceptable based on analysis similar to higher education case law regarding affirmative action. These arguments ignore the fact that many other immutable factors are used in this process, that the stakes are not simply one's educational opportunity but one's liberty, and that much of the higher education case law focuses on remedying historical oppression through measures such as affirmative action. *See* Oleson, *supra* note 8, at 1338.

170 United States v. Virginia, 518 U.S. 515, 533 (1996) (internal quotation marks omitted).

171 *Id.* at 532–34.

172 *See* Starr, *supra* note 1, at 825–26 (noting that cases, such as *Craig v. Boren*, *Weinberger v. Wiesenfeld*, and *Frontiero v. Richardson* showed that strong statistical data between different sexes did not convince the Court of the constitutionality of classifying by sex based on stereotyping). In *Craig*, men were ten times more likely to be arrested for drunk driving than women. Despite this finding, the Court ruled that the drinking law's distinction based on sex was unconstitutional. Craig v. Boren, 429 U.S. 190, 197 (1976). In *Weinberger* and *Frontiero*, women were much more likely to be financially dependent on their husbands (in the 1970s) than the other way around. Despite this finding, the Court ruled that a financially-dependent man was entitled to his wife's benefits. *See* Weinberger v. Wiesenfeld, 420 U.S. 636, 645 (1975); Frontiero v. Richardson, 411 U.S. 677, 690–91 (1973).

173 "Age was the most important demographic factor predicting arrest." IMPACT OF REMOVING DEMOGRAPHIC FACTORS, *supra* note 83, at 1. It is unclear whether this refers to a first or second arrest.

recidivism decreases by roughly 5%.[174] Within the Pennsylvania Commission on Sentencing's nine OGS, age accounts for 21–30% of the risk model's prediction of recidivism.[175] In evaluating whether it could remove this demographic factor from the analysis, the Pennsylvania Commission on Sentencing concluded, "[T]he removal of that factor would have the most impact on recidivism prediction and scale accuracy."[176] If age were removed, older offenders would appear to have higher recidivism rates, whereas younger offenders would appear to have lower recidivism rates as compared to the original scale.[177]

In its most recent validation of Risk Assessment Instrument by OGS for All Offenders for the 2004–2006 development sample, defendants under the age of 21 or younger scored three to five points higher than defendants aged 49 or older.[178] Given that the total risk scores range from eight to nineteen depending upon the defendant's OGS category, age carries a significant weight on the severity of one's sentence. The younger a defendant is, the longer their sentence will be.

This result contradicts the Supreme Court's recent jurisprudence surrounding adolescents. Beginning in 2005, the Supreme Court recognized that adolescents are categorically less culpable than adults.[179] Much of the Supreme Court's logic in subsequent decisions involving juveniles was based off of *amici* briefs showing that "developments in psychology and brain science continue to show fundamental differences between juvenile and adult minds."[180] This scientific evidence notes, "Risk taking [including criminal activity] . . . is so pervasive that it 'is statistically aberrant to refrain from such behav-

---

174 FACTORS THAT PREDICT RECIDIVISM, *supra* note 92, at 9. The author finds it highly unlikely that the researchers intended to convey a linear effect, particularly given that age is grouped into categories.

175 IMPACT OF REMOVING DEMOGRAPHIC FACTORS, *supra* note 83, at 5.

176 *Id.* at 19.

177 *Id.* at 9.

178 VALIDATION OF OFFENSE GRAVITY SCORE, *supra* note 80, at 19.

179 "[P]sychology and brain science continue to show fundamental differences between juvenile and adult minds" making their actions "less likely to be evidence of 'irretrievably depraved character' than are the actions of adults." Miller v. Alabama, 567 U.S. 460, 471–72, 490 (2012) (quoting Roper v. Simmons, 543 U.S. 551, 570 (2005)).

180 Graham v. Florida, 560 U.S. 48, 68 (2010).

ior during adolescence.'"[181] Both the Supreme Court and *amici* focus on young persons' potential for rehabilitation.[182]

Because the Pennsylvania Commission on Sentencing insists on including age in its risk assessment tool, age creates an unfair prejudice. Even though younger people are arguably both legally and scientifically less culpable and more amenable to rehabilitation than their older counterparts, they will likely spend more time in prison.

V
PUBLIC POLICY CONCERNS

Proponents of evidence-based sentencing laud it as the solution for discrimination, overcrowding, and budget shortages.[183] The discussion on its scientific strength above should serve as a warning that evidence-based sentencing may not be the most effective solution.

There are four major penological theories: retribution, deterrence, incapacitation, and rehabilitation.[184] Over time, the purpose of punishment and prison sentences has shifted from rehabilitation, where prisons serve as a way to prepare individuals for return to society, to deterrence and incapacitation.[185] Evidence-based sentencing puts incapacitation at the front and center of the debate. Because recidivism is the focus of evidence-based sentencing, the idea is that the public will be protected while the defendant is imprisoned and thus incapacitated.

Actuarial risk assessments give a simple numerical value—a percentage with questionable accuracy—to answer what appears to the lay person to be the probability that a person will recidivate.[186] This does not provide information about when or how they will recidivate. This snapshot is provided to us after a plea or a trial, but before sentencing, and ignores the many factors that can impact recidivism after and during an individ-

---

181 Brief for the American Medical Association et al., as Amici Curiae Supporting Respondents at 5, Roper v. Simmons, 543 U.S. 551 (2005) (No. 03-633) (quoting L.P. Spear, *The Adolescent Brain and Age-Related Behavioral Manifestations*, 24 NEUROSCIENCE & BEHAV. REVS. 417, 421 (2000)).

182 *See Miller*, 567 U.S. at 483 (noting that mandatory punishment "disregards the possibility of rehabilitation even when the circumstances [(juvenile brain development and culpability)] most suggest it").

183 *See* Starr, *supra* note 1, at 816.

184 *See* Oleson, *supra* note 8, at 1330–32.

185 *See* Michelle S. Phelps, *Rehabilitation in the Punitive Era: The Gap Between Rhetoric and Reality in U.S. Prison Programs*, 45 L. & SOC'Y REV. 33, 34 (2011).

186 Indeed, what qualifies as recidivism is hard to define. *See supra* subpart III.B (Prong I: Whether Evidence-Based Sentencing Has been Tested?).

ual's incarceration (including family ties[187] and prison education programs[188]). Using a single percentage generated by a risk assessment tool does not reduce recidivism; rather, it provides a false feeling of creating a "just" criminal justice system, while failing to rehabilitate and help offenders transition back into society.

CONCLUSION

Despite its purported positive impact on the criminal justice system, evidence-based sentencing risks fooling judges and juries and further contributing to the overrepresentation of men of color and poor people in prisons. The problems with the creation of these models, namely a lack of replication, potentially unconstitutional use of certain factors, high false positive rates, and issues with G2i abstraction, should all create serious concerns for actors in and around the criminal justice system.

The *Daubert* test offers an analytical framework through which the validity and fit of evidence-based sentencing can be evaluated. As a general matter, evidence-based sentencing can and has been tested, but rigor should be applied to sample selection and methods. Peer review should be decided on a case-by-case basis and free from conflicts of interest. The high error rates of specific evidence-based sentencing tools are concerning, particularly with the inability to pinpoint what a large effect size is. Finally, the method of using logistic regression for individual prediction is a generally accepted statistical technique, but the weight to give such information has not been established. The Pennsylvania Commission on Sentencing's current proposal does not pass the first three prongs of *Daubert*, and general acceptance could be argued either way depending on the scope of analysis. Even if Pennsylvania's evidence-based sentencing regime did pass the *Daubert* test, the prejudicial nature of using sex and age would disallow evidence-based sentencing. Whether this sentencing technique would be appropriate without such factors remains an open question.

It will be hard to measure the impact of Pennsylvania's risk assessment tools when they go into effect on July 1, 2018. The two risk assessment tools for a future offense of "Any Crime" or

---

187 Mark T. Berg & Beth M. Huebner, *Reentry and the Ties that Bind: An Examination of Social Ties, Employment, and Recidivism*, 28 JUST. Q., 382, 383 (2011).

188 From class discussion through the Cornell Prison Education Program.

"A Crime Against a Person" may sometimes result in an offender being in different risk groups depending on the test applied.[189] The guidelines offer no explanation on how to reconcile these two outcomes or whether further information is needed if just one of the risk assessments results in a low or high risk. Measuring the outcome of seeking further information for low- or high-risk assessments may be difficult as the judge may order either "a PSI report that contains risk-needs-responsivity information" or "the preparation of an RNA or RNR assessment."[190] After considering its methodological and constitutional shortcomings, it is unclear whether Pennsylvania's new risk assessment tools will truly result in "smarter sentencing," and whether we will know what it did at all.

189 PROPOSALS PUBLISHED, *supra* note 85, § 305.9, Risk Assessment Summary (the Commission failed to label § 305.9 in the original report).
190 *Id.* § 305.5(c).